INF2178_A4

Student Name: Nianchuer Liu

Student Number: 1010332454

Email:nianchuer.liu@mail.utoronto.ca

a. Preliminary Analysis and Research Questions

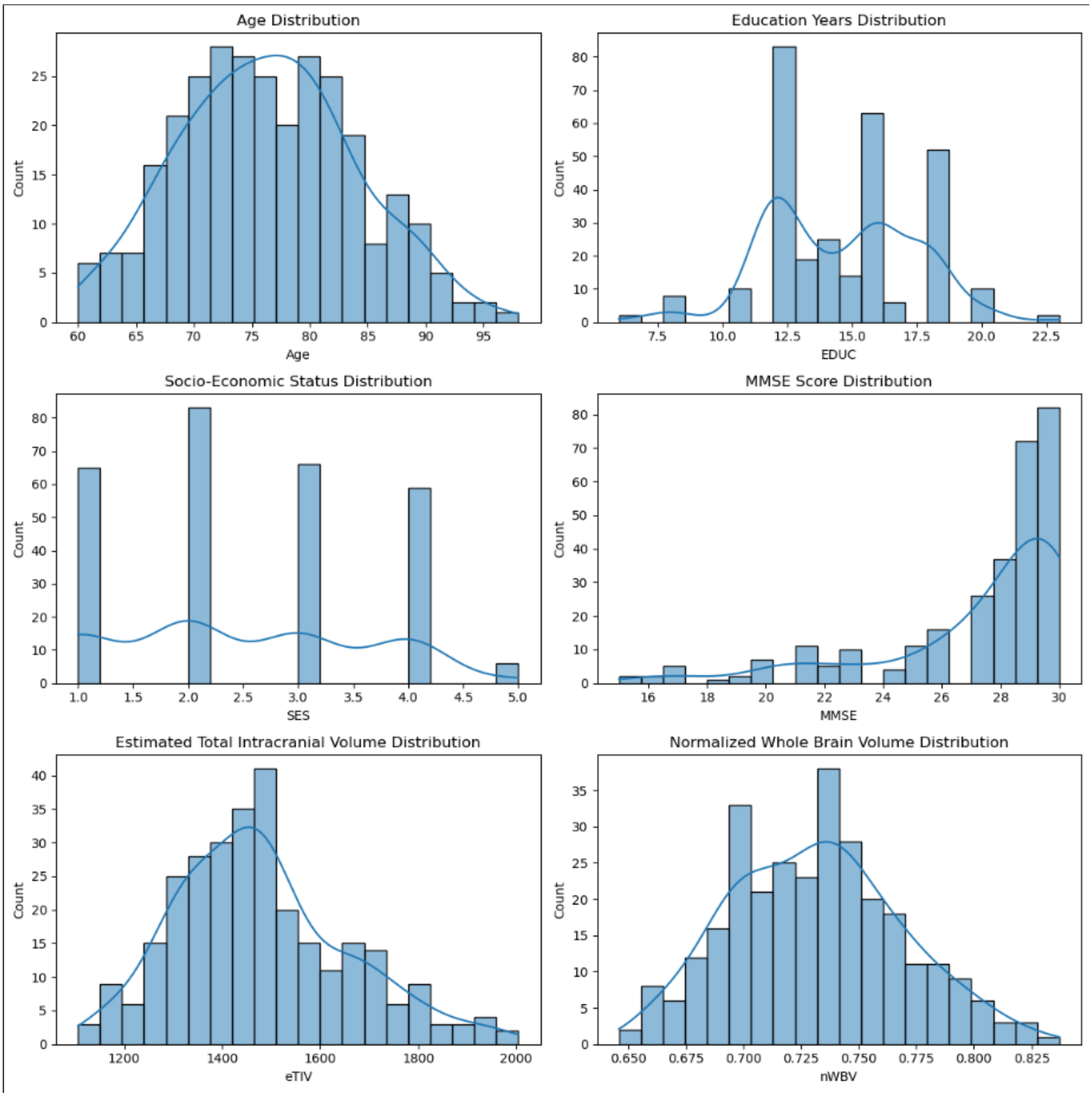
There are many columns in the dataset, such as an unnamed column that is probably an index, Subject ID, MRI ID, Group (which sorts subjects into "Nondemented" or "Demented"), Visit number, MR Delay (which is likely the number of months since the first visit), gender (M/F), hand preference (Hand), age, education level (EDUC), socio-economic status (SES), Mini-Mental State Examination score (MMSE), Clinical Dementia Rating (CDR), estimated total intracranial volume (eTIV), normalize, and

Based on a first look, this dataset seems to be linked to a study on dementia. It may be looking into how age, education, socioeconomic status, and brain volume measurements affect the status of dementia. The fact that there were both "Nondemented" and "Demented" groups, and that some people were seen more than once, suggests that the participants were followed over time to see how they changed or how they got dementia.

Given the dataset's focus on dementia, a key research question could be: "How do demographic and neuroimaging metrics influence the likelihood and progression of dementia over time?" Specifically, we could explore:

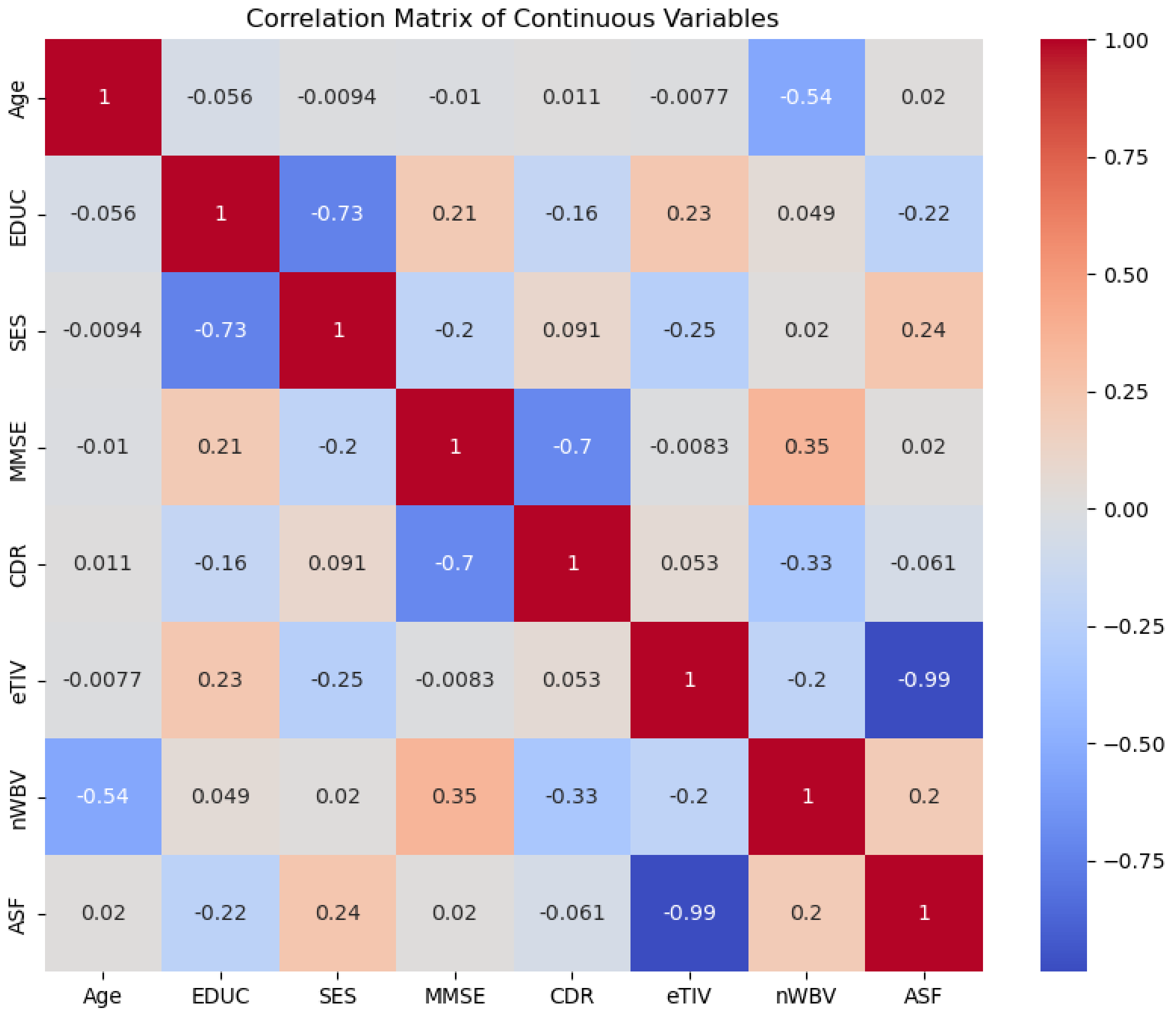
1. The effect of age, education, and socio-economic status (SES) on dementia status.
2. The relationship between brain volume measurements (eTIV, nWBV) and dementia progression

b. Exploratory Data Analysis (EDA)



1. **Age Distribution**: The distribution of age appears to be slightly right-skewed, indicating a larger proportion of older individuals in the dataset.
2. **Education Years Distribution**: The number of years of schooling seems to be spread out in a multimodal way, with high points around 12 years (high school graduation) and 16 years (college graduation). This shows that the participants had different levels of education, which is normal for a diverse group of people.
3. **Socio-Economic Status (SES) Distribution**: The SES scores are skewed to the right, which means that most of the subjects have lower SES levels. The skewness could be a sign of underlying differences in socioeconomic status or trends of hiring.
4. **MMSE Score Distribution**: The Mini-Mental State Examination (MMSE) scores tend to lean to the left, with most people getting higher, which means they have better cognitive function, but there is a tail of scores that are lower. This trend fits with a group of people who have both normal cognitive functioning and different levels of cognitive impairment.
5. **Estimated Total Intracranial Volume (eTIV) Distribution**: The eTIV distribution looks pretty normal, which means that the total intracranial volume changes within a normal range in the group that was studied. It is likely that this normal distribution will hold true since eTIV shouldn't be directly affected by disease state. Instead, it should be affected by differences between people's head sizes and brain development.

6. **Normalized Whole Brain Volume (nWBV) Distribution**: The distribution of nWBV is also pretty normal, but there is a small leftward shift that points to a general trend toward smaller brain sizes in the population. This might be because the study included people with diseases like dementia, which are linked to brain atrophy.



The correlation matrix shows how the continuous variables in the dataset are related to each other. This can help us with our research and model choice. For instance, there is a strong negative relationship between "Age" and "nWBV" (Normalized Whole Brain Volume), which suggests that brain volume may drop with age. This is in line with what researchers have found in studies about dementia and getting older.

c. Mixed-effects ANOVA models

Model 1: Predicting MMSE Score

- **Dependent Variable**: MMSE (Mini-Mental State Examination score)
- **Predictors**: Age, EDUC (Years of Education), and Group (Demented vs. Nondemented)
- **Random Effect**: Subject ID

Result:

1. The condition of the group strongly predicts MMSE scores; people who were demented scored, on average, 3.879 points lower than the baseline group ($p < 0.001$).
2. In this model, age and schooling were not significant predictors. This suggests that dementia status has a more direct effect on cognitive scores as measured by the MMSE than these demographic factors.

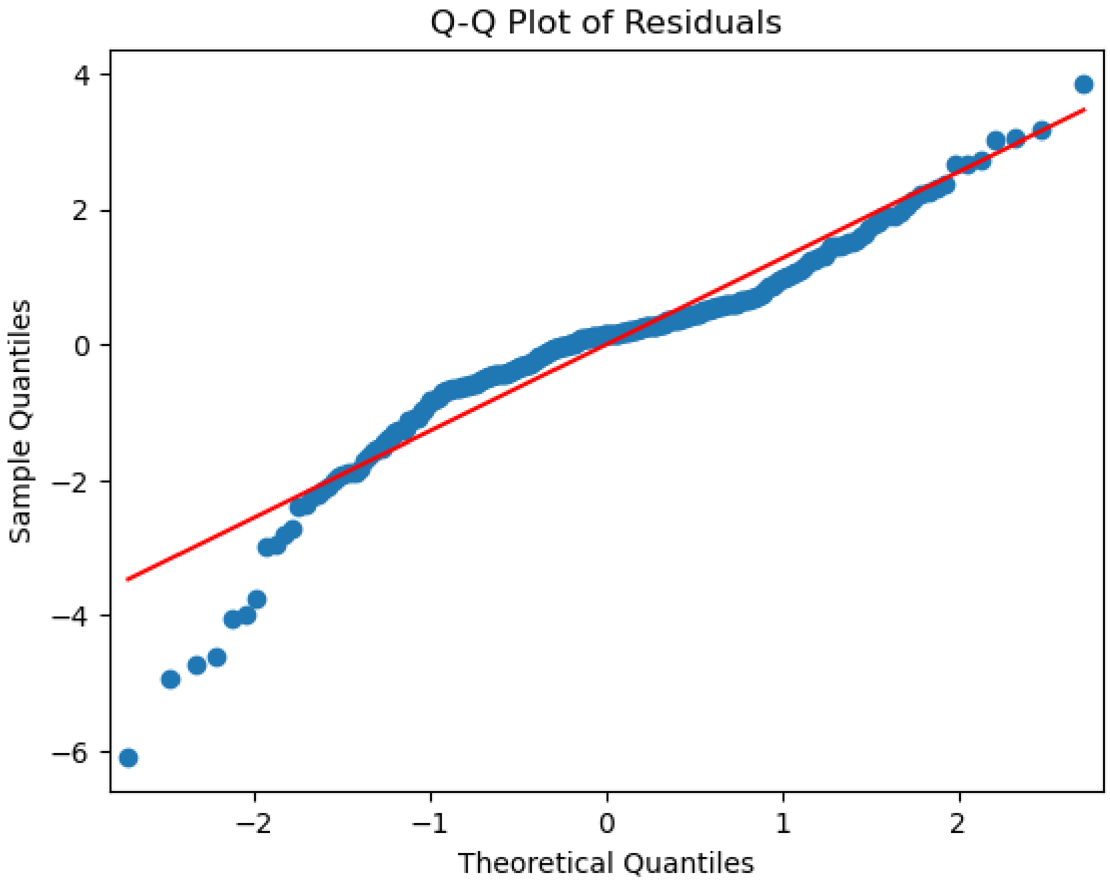
Model 2: Predicting Normalized Whole Brain Volume (nWBV)

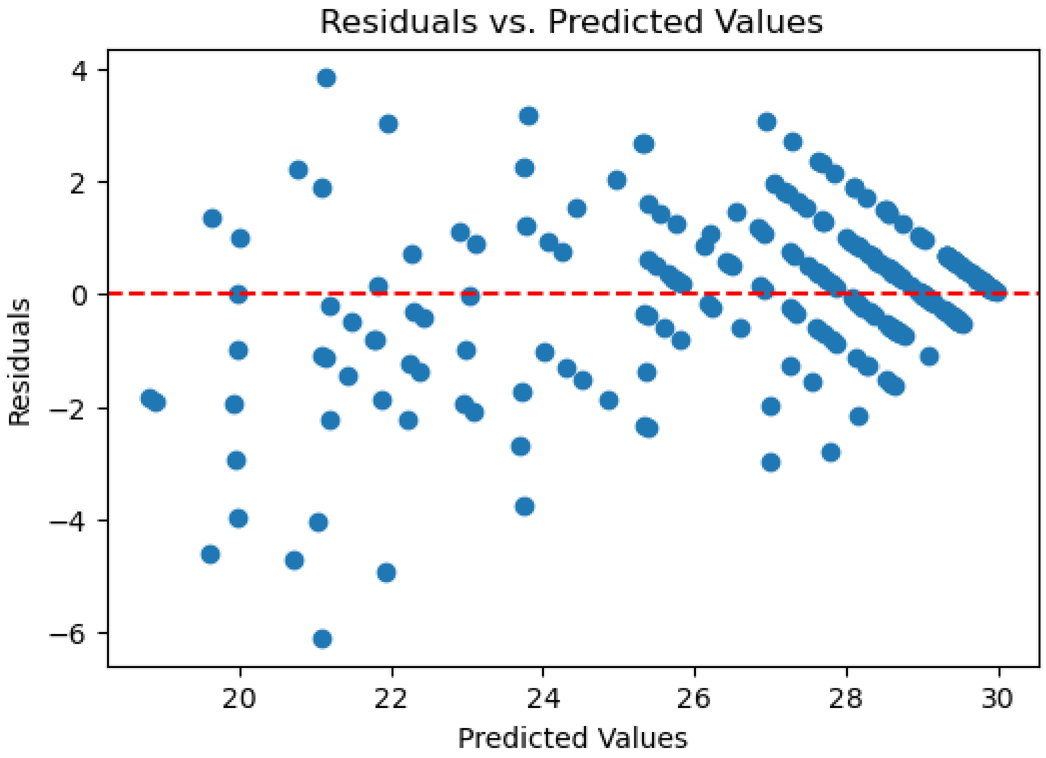
- **Dependent Variable**: nWBV
- **Predictors**: Age, EDUC, and Group

Result:

1. There is a strong link between getting older and less brain volume, as shown by the fact that nWBV goes down by 0.003 per year of age ($p < 0.001$).
2. The fact that dementia is linked to a drop in nWBV (coefficient = -0.022, $p = 0.011$) shows how dementia changes brain volume.
3. In this model, education level and not having dementia did not significantly predict nWBV.

d. Testing the assumptions





Normality of Residuals

- **Shapiro-Wilk Test**: The test statistic is approximately 0.922, and the p-value is significantly low ($p < 0.05$), suggesting that the residuals do not follow a normal distribution. This could indicate that the model's assumptions about the distribution of residuals are violated.
- **Q-Q Plot**: The Q-Q plot visually supports this finding, as the points deviate from the straight line, especially in the tails, indicating non-normality in the distribution of residuals.

Homogeneity of Variance (Homoscedasticity)

- **Residuals vs. Predicted Values Plot**: This graph is used to check if the residuals have the same amount of variation across the expected value range. In an ideal world, the spread of residuals would be about the same at all expected value levels, with no clear pattern. There is some spread in the plot, but there isn't a clear trend of rising or falling variance across the range of predictions. This is a good sign for homoscedasticity. But the fact that the residuals are not normal could change this view.

e. Power Analysis

For a theoretical experiment aiming for a power of 0.91, an alpha level of 0.05, and an effect size of 0.7, the appropriate sample size needed per group is approximately 46 participants. This means that to detect an effect of this size with the specified power and significance level, each group in the study should consist of at least 46 subjects.